 **System Requirements:** Make sure you are using Ubuntu and have Python 3.7 or higher installed. Additionally, ensure you have a C++ compiler supporting C++17 and OpenMP, and if you are using a GPU, make sure the NVIDIA drivers and CUDA toolkit are properly installed.

 **Clone the Repository:** First, clone the ParEval repository and make sure you initialize all submodules. Go into the project folder once it's cloned.

 **Install Python Dependencies:** Install all required Python dependencies by referring to the requirements.txt file included in the repository. If you are working with models that use a GPU, ensure your system's GPU drivers and CUDA toolkit are set up correctly.

 **Build Kokkos (Optional):** If you want to include Kokkos in your tests, follow the build instructions provided in the repository. You'll need to use CMake to configure and build it.

 **Prepare Prompt Files:** Ensure that the prompt file generation-prompts.json is placed in the prompts directory. This file is already included in the repository and will be used to generate model outputs.

 **Choose a Suitable Model:** Depending on the hardware available, choose an appropriate model. For lower memory systems, you can select smaller models like PolyCoder (2.7B parameters) or Replit (3B parameters). The list of supported models can be found in the utils.py file.

 **Run the Generation Script:** Use the generate.py script to generate outputs from the model. You'll need to provide the model name, the prompt file, and the output file path. For example, you can run the script with the PolyCoder model or Replit model, depending on your setup. Ensure that you configure caching appropriately by providing a cache file.